Stable equilibria and forward induction

van Damme, E.E.C.

Published in:
Journal of Economic Theory

Publication date:
1989

Reprinted from JOURNAL OF ECONOMIC THEORY
Vol. 48, No. 2, August 1989
Printed in Belgium

Stable Equilibria and Forward Induction*

ERIC VAN DAMME†

Universität Bonn, Adenauerallee 24-26,
D-5300 Bonn 1, West Germany

Received October 16, 1987; revised June 1, 1988

This paper is an attempt to throw some light on the issues of whether requiring an equilibrium to be stable (in the sense of Kohlberg and Mertens) is necessary for self-enforcingness and what the implications of such a requirement are. In the first part it is discussed which role "mistakes" play in the stability concept and it is argued that stability does not fully capture the logic of forward induction. The second half is devoted to specific examples that show the power of stability and that indicate that a large part of noncooperative game theory may have to be modified in an essential way if one accepts stable equilibrium as the solution concept. *Journal of Economic Literature* Classification Number: 026.

1. INTRODUCTION

It has long been recognized that not every Nash equilibrium is self-enforcing, i.e., not every equilibrium has the property that, when it is recommended to the players, no player has an incentive to deviate from it. In extensive form games, this phenomenon is caused by the fact that some equilibria are sustained only by incredible threats off the equilibrium path. Players moving on the equilibrium path will realize that these threats are empty, hence, they may be inclined to deviate. The concepts of sequential and (subgame) perfect equilibria (cf. Kreps and Wilson [11] and Selten [18]) have been proposed to resolve this problem but it has been realized that (at best) these only yield necessary conditions for strategic stability; they do not impose sufficient restrictions on out-of-equilibrium beliefs so that a path may be sustained only by incredible beliefs. Basically the two concepts mentioned do not provide a solution since they allow out-of-equi-

* Parts of this paper were presented at seminars and conferences at Northwestern University (Aug. 86), Bielefeld (Oct. 86 and 87), Paris (Dec. 86), Luminy (Jan. 87), Bonn (June 87), and the LSE (Feb. 88). The author thanks the participants for their comments and suggestions. A special thanks goes to Martin Hellwig and Reinhard Selten for many stimulating conversations and to Jean-Francois Mertens for extremely helpfull correspondence. The usual disclaimer applies.

† Supported by the Sonderforschungsbereich 303 (DFG), Universität Bonn, W. Germany.

0022-0531/89 $3.00

librium moves to be ignored as they may have occured by mistake. Non-self-enforcing equilibria may also exist in normal form games because of two reasons: (i) equilibria may vanish when (weakly) dominated actions are deleted from the game (hence such equilibria are not viable if players adhere to a theory that tells that dominated strategies are irrational) and (ii) equilibria need not be robust, so that the slightest uncertainty about the motives or the rationality of the opponents may lead players away from equilibrium.

Recently, Elon Kohlberg and Jean-François Mertens [9] (henceforth KM) proposed the concept of stable equilibrium to eliminate those equilibria suffering from any of the three drawbacks mentioned above. Some simple examples and preliminary applications have made it clear that "stability" is fundamentally different and much stronger than any other existing refined equilibrium notion. However, as the KM approach is rather abstract and indirect (by requiring robustness in the normal form it aims at obtaining both forward and backward induction in the extensive form), it has remained somewhat obscure whether stability is a necessary (and/or sufficient) condition for self-enforcingness and what its implications really are. This paper is an attempt to shed some light on these issues. In particular, in Section 2 it is investigated whether normal form analysis is appropriate and it is also demonstrated that perturbations (mistakes) play a fundamentally different role in the KM theory as they do in Selten's trembling hand perfectness concept. In Section 3 the connections between stability and forward induction are explored and it is argued that not every stable equilibrium is consistent with a forward induction logic. Sections 4 and 5 are devoted to specific examples illustrating the power of stability and the surprising strength of requiring a solution to be "invariant" with respect to elimination of dominated strategies. These examples clearly indicate that a large part of noncooperative game theory will have to be modified in an essential way if one accepts stable equilibrium rather than Nash (or sequential/perfect) equilibrium as the relevant solution concept.

Before starting the discussion it is worthwhile to draw attention to Footnote 3 of KM. There it is made clear that stability is a pure noncooperative solution concept; i.e., it requires that all aspects relevant to the situation be explicitly modeled by the rules of the game. This is the more important as it will turn out that stable equilibria are very sensitive to modeling "details." Hence, stability is inconsistent with "small worlds" arguments and if one cannot (or does not wish to) model all details, then stability should not be used as the solution concept.

2. STABLE EQUILIBRIA

Loosely speaking, an equilibrium outcome (path) of a (generic) extensive form game is stable if the set of normal form equilibrium strategies that sustain this path is stable with respect to arbitrary slight perturbations in strategies, i.e., if for any slightly perturbed normal form there exists an equilibrium close to the original set.[1] Note that stability is a normal form solution concept and that stability does not refer to single equilibria but rather to equilibrium paths (so to sets, or better, components, of equilibria). In addition to establishing general existence of stable sets (in fact, stable components) and generic existence of stable paths, KM prove two important properties which we will use extensively when analysing our examples:

PROPOSITION A (Kohlberg and Mertens [9]). (i) *A stable set contains a stable set of any game obtained by deletion of a dominated (pure) strategy.*

(ii) *A stable set contains a stable set of any game obtained by deletion of a strategy that is an inferior response against the set, i.e., that is not a best response against any element of the set.*

Since the formal definition of "stability" involves perturbations, stability superficially resembles Selten's [18] trembling hand perfectness concept. There are, however, two main differences:

(i) Stability is a normal form solution concept, whereas Selten's perfectness notion is defined by means of the agent normal form. Hence, perfectness assumes independent mistakes at different decision points. Normal form considerations introduce a kind of correlation between different "mistakes."

(ii) Perfectness is a single-valued solution concept, whereas for stability it is essential to consider sets of equilibria. Stability only describes which outcomes can be expected to be realized by "rational" players, it does not always tell exactly what a player should do off the equilibrium path.

We will comment on these aspects below.

At the end of the section we will return to the second issue, let us now discuss the first point of difference. At the intuitive level this point corresponds to saying that perfectness too soon allows players to conclude that a mistake has been made; stability requires each player first to scrutinize the possibility that behaviour was intended and rational after all.

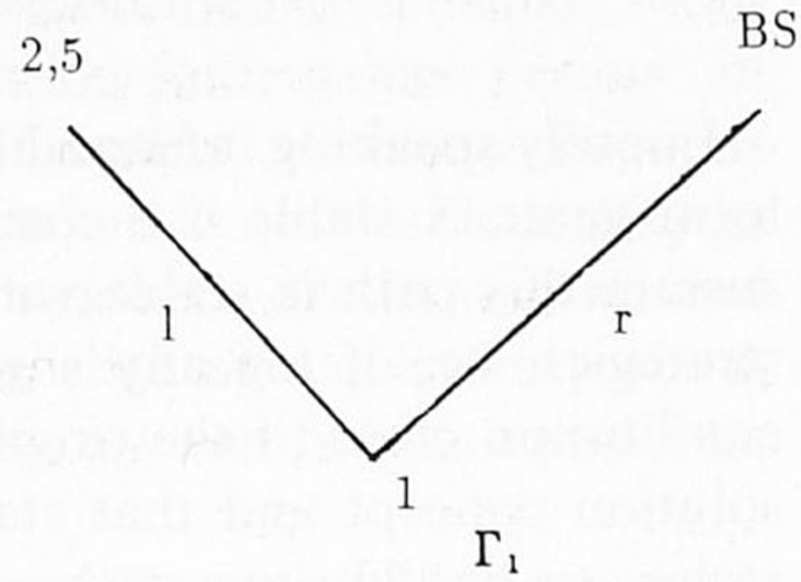
[1] For a formal definition see the Appendix in which all other technical terms that are used in the paper are defined also.

a

	s	w
s	0, 0	3, 1
w	1, 3	0, 0

BS

b



Γ_1

c

	s	w
l	2, 5	2, 5
rs	0, 0	3, 1
rw	1, 3	0, 0

G_1

FIG. 1. The battle of the sexes: (a) BS; (b) Γ_1; (c) G_1.

The game of Fig. 1 may illustrate the difference. Figure 1a is the battle of the sexes (BS) which will occur as an essential ingredient in many of our examples. In Fig. 1b, player 1 has to decide to take his outside option l (with payoff (2, 5)) or to play BS; when player 2 has to move, he knows that 1 has not chosen l. Figure 1c is the reduced normal form of Fig. 1b.

Assume players have agreed that 1 will play l in Γ_1 and that they will continue with (w, s) in the unlikely event that BS has to be played. This equilibrium is perfect: If player 2 is reached, this player could conclude that 1 suffered from "momentary insanity" at his first decision point, but he could still presume that this player will play rational (i.e., continue with the plan) at the second node, and in this scenario s is indeed the unique best response of player 2. The equilibrium, however, is not stable: Player 2 should not rush to conclude that 1 has made a mistake, he should realize that a fully rational player will never play a dominated strategy, (in this case rw); hence, he should conclude that 1 has planned rs and he should

respond with w which upsets the equilibrium.[2] (Formally, iterated elimination of (weakly) dominated strategies reduces the normal form G_1 to (3, 1) so that Proposition A shows that the equilibrium component (2, 5) cannot be stable.)

Note that the above argument for instability of "(2, 5)" in fact excludes the possibility that player 1 made a mistake[3]; hence, there is a fundamental difference between the perturbations in Selten's theory and those in KM's. Indeed, Selten writes:

> There cannot be any mistakes if the players are absolutely rational. Nevertheless, a satisfactory interpretation of equilibrium points in extensive games seems to require that the possibility of mistakes is not completely excluded (Selten [18, Sect. 7]),

and this is in sharp contrast to KM's statement that

> probabilities of error—such as those occurring in the definition of "perfect equilibrium" must not be interpreted as probabilities that the players will actually err in choosing their strategies (footnote 3, *in loc cit*).

KM have good reasons for taking this position, for, if independent mistakes are a real possibility, stability is not the correct concept. In this case the perturbations should not be included in the normal form but in the extensive form, since the normal form restricts the mistakes a player can make (one can only make them in the beginning) and prevents a player from correcting his mistakes. Hence, if independent mistakes are real, one should interchange the two operations (normalisation and perturbation) in the KM procedure. The game of Fig. 1b illustrates that this makes a crucial difference. Assume (for simplicity only) that the only mistakes that can occur in Γ_1 (Fig. 1b) is that player 1 chooses r and that this mistake occurs with probability ε. In this case ls and lw are not equivalent strategies in the normal form. This normal form is now given in Fig. 2 and although rw is

[2] Note that the argument for instability of (2, 5) would remain unchanged if the payoffs in BS associated with (s, w) and (w, s) were $(2+\varepsilon, 1)$ and $(1, 3/\varepsilon)$ respectively with $\varepsilon > 0$, even though now 2 presumably has stronger arguments for pushing (w, s) if BS were played in isolation. However, BS is not played in isolation so that 2 knows that the payoff $3/\varepsilon$ is illusory (hence, irrelevant) as it can only be realised if 1 plays a dominated action which he (being fully rational) will never do.

[3] The argument also breaks down if player 2 is slightly uncertain about 1's payoffs. In particular, if there is a small probability ε that player 1's payoff associated with (w, s) in BS is more than 2, then when player 2 is reached he can always believe that this possibility prevails, so that 1 will be forced to choose l in case his payoffs are actually as in Γ_1. Hence, (l, s) can be approximated by strict (hence stable) equilibria of "nearby" games with incomplete information (see Fudenberg, Kreps, and Levine [10] for the proof that this construction works for almost any pure Nash equilibrium).

	s	w
ls	$2-2\varepsilon, 5-5\varepsilon$	$2+\varepsilon, 5-4\varepsilon$
lw	$2-\varepsilon, 5-2\varepsilon$	$2-2\varepsilon, 5-5\varepsilon$
rs	0, 0	3, 1
rw	1, 3	0, 0

$G_1(\epsilon)$

FIG. 2. The normal form, $G_1(\varepsilon)$.

still (strictly) dominated for player 1, s is no longer dominated for player 2 (for $\varepsilon > 0$) as this player now can no longer be sure that player 1 has not made a mistake. In fact, for $\varepsilon > 0$ one has that (lw, s) is a strict (hence, stable) equilibrium of $G_1(\varepsilon)$ and this corresponds to the perfect (unstable) equilibrium with which we started our discussion of Γ_1.

From the above example,[4] we may conclude that there is an essential difference between Selten's perfectness concept in which independent mistakes are real (hence, complete rationality is viewed as a limiting case of incomplete rationality) and KM stability which considers mistakes an irrelevant possibility (or at least, when mistakes are relevant they should be explicitly modeled). Hence, KM insist that the players are perfectly rational, and this of course raises the issue of how the perturbations in the definition of stability should be interpreted. In this author's view, the KM position implies that the current definition should be viewed merely as a computational technique for checking self-enforcingness; it should be possible to define stability just in terms of the decision theoretic structure of the game, i.e., without using trembles. (Mertens [12] has indicated that this should be possible (hence, that the heuristic argument from Appendix D of KM can be converted into a theorem), but unfortunately it is still unknown

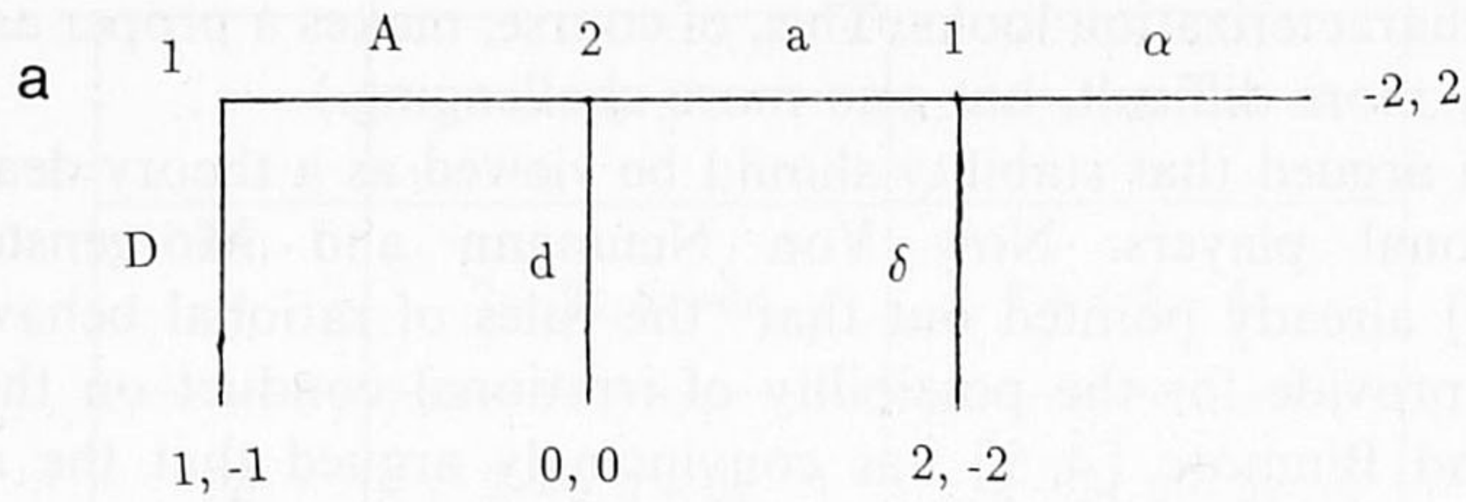
[4] The example is in no way special: The same construction can be done for (almost) any pure perfect equilibrium: Perturbing first will convert it into a strict equilibrium. It should also be remarked that the point we make is related to the fact that the equilibrium outcome "2, 5" is stable in the agent normal form: The iterated dominance fails if the decisions at the 2 nodes of player 1 are made fully independently. Any perturbed game associated with the agent normal form has an equilibrium close to (l, w, s).

how this characterization looks. This, of course, makes a proper assessment of stability more difficult, but also more challenging.)

Above I argued that stability should be viewed as a theory dealing with fully rational players. Now Von Neumann and Morgenstern [20, Sect. 4.1.2] already pointed out that "the rules of rational behavior must definitely provide for the possibility of irrational conduct on the part of others" and Binmore [4, 5] has convincingly argued that the notion of perfect rationality may not even be meaningfull (as opposed to Selten's notion of the limit of incomplete rationality). The point is that in planning his decisions a player assumes that the players are rational (i.e., behave in accordance with a certain theory) but that some information sets may be reached only if this assumption is violated. Hence, the question is what a theory of perfect rationality should prescribe at such an information set and, more generally, whether there can exist a logically consistent theory of perfect rationality that does not involve the idea of mistakes. KM circumvent (or answer?) these questions by conceding that a theory of perfect rationality cannot be single-valued; i.e., after a player has taken an action that is not in accordance with the theory one should not automatically assume that a player will behave rationally later on in the game. In this respect the theory again differs from perfectness in which, as noted above, an assumption of "momentary insanity" is made. Consequently, even though the number of stable equilibrium paths is usually smaller than the number of perfect equilibrium paths, stability may in fact allow a greater freedom in off the equilibrium path behaviour. The zero-sum game of Fig. 3 provides an example.

Player 1 should play D in any Nash equilibrium of the game of Fig. 3 so that player 2 has to move only when 1 has not acted in accordance to conventional game theory. What should 2 conclude? According to (subgame) perfectness, 2 should believe that 1 suffered from temporary insanity but that he will nevertheless play rational (i.e., choose δ) later. But why should player 2 believe this; could not he bet on player 1 being fully irrational and play a instead of his subgame perfect equilibrium strategy d? The problem, of course, is that a rational player 1 would gain by choosing A if thereby he could make player 2 believe that he is irrational. Nevertheless, as Binmore [5, p. 23] has argued, the restrictive assumption of player 2 that his opponent is irrational with probability zero seems unjustified. In fact, "stability" does not make this assumption: The subgame perfect equilibrium (D, d, δ), taken as a singleton is not stable, the unique stable set includes both (D, d) and $(D, \frac{1}{2}a + \frac{1}{2}d)$. Hence, intuitively, stability does not force 2 to believe that 1 will play δ, player 1 may be irrational but 2 should not make it attractive for a rational player to pretend to be irrational. Of course, this raises the question of which behavioural assumption justifies playing $\frac{1}{2}a + \frac{1}{2}d$ for player 2? This example also shows that a stable set may

a



b

	d	a
D	$1,-1$	$1,-1$
δ	$0,0$	$2,-2$
α	$0,0$	$-2,2$

FIG. 3. Zero-sum game.

contain more outcomes than the ones that survive iterated elimination of dominated strategies (hence, it is essential that one has "contain" in Proposition A and not "is"), and the foregoing discussion implies that this is for good reasons. Namely, even though it is irrational to play a dominated strategy (cf. Sect. 2.7 of KM), the procedure of iterative elimination of such strategies is justified only if all players are fully rational. Intuitively, stable sets have to be large as they must incorporate the possibility of irrational play (and there seems no unique optimal way to play against irrational opponents); however, as soon as one starts eliminating dominated strategies of the opponents, one attributes more rationality to them, hence, makes them more predictable and this leads to a smaller set of optimal actions, hence, to smaller stable sets.

3. FORWARD INDUCTION

KM argue that essentially what is involved in the discussion of game Γ_1 (Fig. 1b) is an argument of "forward induction." When player 2 hears that he has to move in Γ_1 he knows that player 1 is effectively sending the following message:

> Look, I had the opportunity to get 2 for sure and nervertheless I decided to play this subgame and my move is already made. And we both know that you can longer talk to me because we are **in** the game. So think now well and make your decision" (p. 1013, *in loc cit*)[5]

In Γ_1 it is completely clear what player 2 should think (as *rw* is dominated) but in more complicated games there may be ambiguity so that one can imagine various formalisations of the idea of forward induction. In this section, a formalisation will be introduced that is based on Schelling's [17] focal point idea and it will be shown that stability does not satisfy this version of forward induction.[6] In this author's opinion, the example (Fig. 4) shows that not every stable equilibrium can be considered self-enforcing.

To illustrate the idea, consider again game Γ_1 which contains BS as a subgame. Furthermore, assume players consider either Nash, perfect, sequential, or stable equilibrium to be the relevant solution concept. All these concepts accept three solutions for BS, viz. (s, w), (w, s), and (m, m) where $m = s/4 + 3w/4$; hence, players realize that they always should coordinate on any of these. However, as soon as BS is embedded in Γ_1 only the solution (s, w) is focal. Player 2 should ask himself according to which solution player 1 will play given that this player has not chosen his outside option l. Given that only (s, w) yields player 1 more than l does, the only sensible conclusion can be that 1 will play s and player 2 can do nothing but go along. (If player 1 indeed chooses s then player 2's unique best response is w.) Hence, if 1 does not choose l, then he credibly signals that he will play according to (s, w) in BS and, therefore, only (rs, w) is self-enforcing in Γ_1. Note that the essential difference between this forward induction concept and that of KM is that the above does not require player 2 to scrutinize which strategy player 1 might play but rather it requires player 2 to investigate according to which solution player 1 might play in the subgame. So in a sense the current approach assumes more rationality on the part of the players than do KM.

In this paper I will not give a formal definition of forward induction[7] but

[5] As Martin Hellwig pointed out to me, this argument actually goes fully against the spirit of the KM paper. If the normal form is relevant, one is always in the game and moves are simultaneous.

[6] Actually, KM use the term "forward induction" for property (ii) of Proposition A, but this is misleading. Namely, if in Γ_1 the payoff (0, 0) associated with (rw, w) is replaced by a zero-sum subgame with value 0 in which player 1 can get more than 3 (if the opponent does not play optimally), then Proposition A does not allow reduction of the reduced normal form, even though the strategic situation is unchanged. Hence, one needs the full power of stability to get solutions with a forward induction flavor. (Indeed only (3, 1) is stable in the modified game.)

[7] For this see the original discussion paper version van Damme [19]. Related ideas occur in Ponssard [16] and Weibull [21].

rather I will state a (weak) property which in my opinion should be satisfied by any concept that is consistent with forward induction. Yet, I will show that stability does not satisfy this property.[8] The proposed requirement is that in generic 2-person games in which player i chooses between an outside option or to play a game Γ of which a unique (viable) equilibrium e^* yields this player more than the outside option, only the outcome in which i chooses Γ and e^* is played in Γ is plausible.

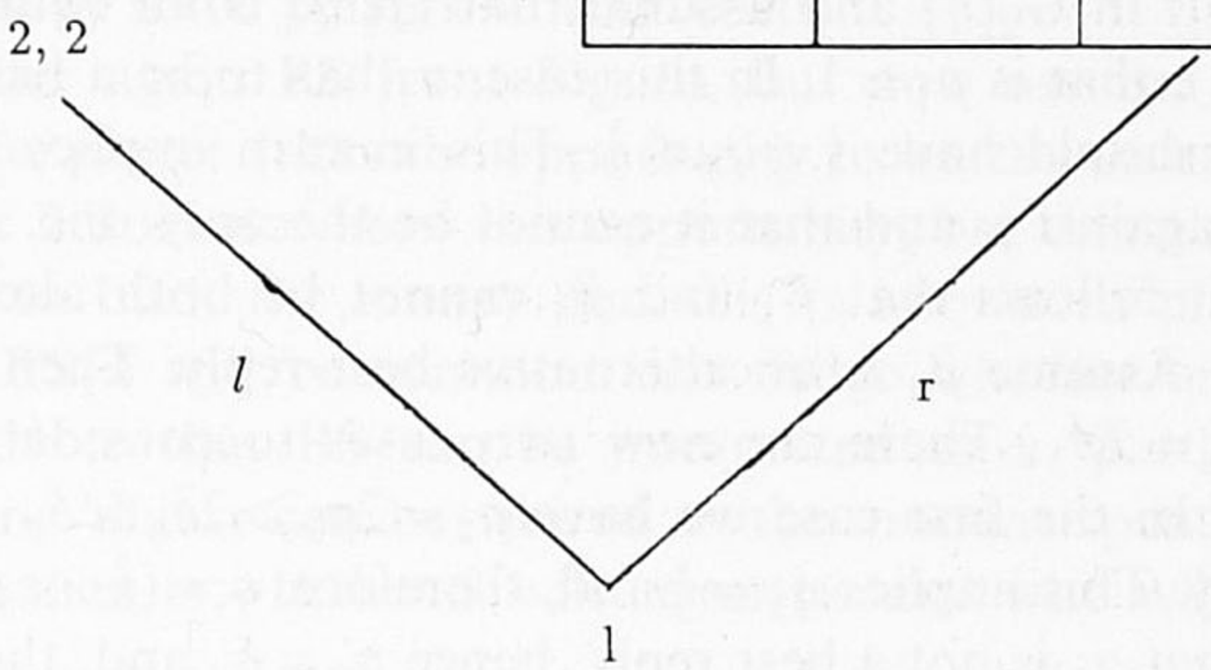
Before turning to the example let me briefly address the qualifications that e^* be unique and viable. The need for them is best illustrated by example. If, in BS both (s, w) and (w, s) would yield player 1 the payoff 3, then if player 2 had to move in Γ_1 he could not know the intentions of player 1 and player 2 would be justified in playing $1/2s + 1/2w$ in which case it is optimal for player 1 to choose his outside option. Hence, it is important that player 1 can unambiguously signal his intentions, therefore, uniqueness was required. To illustrate the need for viability of e^*, consider the game in which player 2 chooses between l, an outside option with payoff (4, 4), or r, that is, to play the game Γ_1 from Fig. 1b. Now, Γ_1 has two Nash (or sequential) equilibrium payoffs, viz. (3, 1) and (2, 5) so that one might argue that, by choosing r, player 2 signals that he wants the payoff (2, 5) (as this is the only equilibrium continuation that yields the player more than his outside option), implicitly threatening with the signal that he will play (w, s) (or the mixed equilibrium) in BS. If player 1 believes this threat then he should play l, thereby indeed granting player 2 the payoff 5. In my opinion, player 2's threat is empty and player 1 should play rs. Namely, above we have seen that only the payoff (3, 1) is viable in Γ_1 (i.e., at least if the players accept forward induction) so that a fully rational player 2 will realize that if Γ_1 is reached he will only obtain 1, hence player 2 should choose l and the only viable payoff in Γ_2 is (4, 4). (Note that this is indeed the only payoff that survives iterated elimination of dominated strategies in the reduced normal form.) Hence, forward induction can only determine which solution should be played in a subgame, but a solution of the game should always induce a solution in each subgame; backwards induction ranks above forward induction.

The game Γ_2 of Fig. 4 may show that stability need not be consistent with the forward induction logic advanced in this section. (The payoffs may be perturbed so as to make the game generic.) The subgame that occurs in this game has three Nash equilibria, viz. (T, L), $\langle(\frac{2}{3}, \frac{1}{3}), (\frac{1}{2}, \frac{1}{2}, 0)\rangle$ and $\langle(\frac{1}{3}, \frac{2}{3}), (0, \frac{1}{2}, \frac{1}{2})\rangle$. The latter two yield player 1 only $\frac{3}{2}$ (less than 2), whereas the former yields 3. This equilibrium (T, L) is clearly viable in the subgame.

[8] One reaction may be: This is not surprising at all since stable equilibria need not even be consistent with backwards induction (cf. Gul's example in KM). However, that phenomenon seems to be caused only by the fact that one considers stable sets that are not contained in one component. In this paper, attention is restricted to stable components.

a

	L	M	R
T	3, 3	0, 2	3, 0
B	0, 0	3, 2	0, 3



Γ_2

b

	L	M	R
l	2, 2	2, 2	2, 2
rT	3, 3	0, 2	3, 0
rB	0, 0	3, 2	0, 3

G_2

FIG. 4. Not every stable equilibrium can be considered self-enforcing: (a) Γ_2; (b) G_2.

In fact, it is a strict equilibrium so that each player loses if he unilaterally deviates from it. In my view, if player 1 chooses to play the subgame he unambiguously signals that he will play T and, thereby, he forces player 2 to choose L. Consequently, player 1 has an incentive to play the subgame and, in my view, only (rT, L) is self-enforcing in the overall game. This equilibrium is indeed stable, but the point is that it is not the only one[9]:

[9] I have not succeeded in finding an example in which no stable equilibrium is consistent with the forward induction logic, so it may be conjectured that an appropriate refinement of stability has the desired property. Note that (2.2) is *not* fully stable in G_2 (*KM*, p. 1024).

Also the component (2, 2) in which player 1 chooses l with probability 1 is stable in G_2.

To point out the mathematical reason for why stability does not conform with forward induction in this example, I will compute, for every perturbed game, all equilibria close to (2, 2), rather than just showing that (2, 2) is stable. Denote the pure strategies of player 1 (resp. player 2) by α_i (resp. β_i) (hence $\alpha_1 = l$, etc.), let δ_i (ε_i) be the minimum probability that a player has to assign to α_i (β_i) in $G_2(\eta)$ and assume that (p, q) is an equilibrium of $G_2(\eta)$ close to (2, 2), that is $p_1 \approx 1$. In this case, α_1 has to be a best reply against q so that we should have $\frac{1}{3} \leqslant q_2 \leqslant \frac{2}{3}$. This in turn implies that β_2 must be a best reply against p and that it cannot be the only one. Since p is completely mixed it follows that β_1 and β_3 cannot be both alternative best replies against p. Assume β_1 is an alternative best reply. Then $q_3 = \varepsilon_3$ and we must have $p_2 = 2p_3$. There are now two cases to consider, either $\delta_2 < 2\delta_3$ or $\delta_2 > \delta_3/2$. In the first case we have $p_2 = 2p_3 \geqslant 2\delta_3 > \delta_2$ so that α_2 must be a best reply. This implies $q_2 = \frac{1}{3}$ and, therefore, $q = (\frac{2}{3} - \varepsilon_3, \frac{1}{3}, \varepsilon_3)$. Finally, this implies that α_3 is not a best reply, hence $p_3 = \delta_3$ and, therefore, $p = (1 - 3\delta_3, 2\delta_3, \delta_3)$. We have found the first equilibrium of $G_2(\eta)$ close to (2, 2) for the case $\delta_2 < 2\delta_3$. All other cases can be analysed similarly by using that α_1 and β_2 have to be best replies and that α_2 and α_3 (resp. β_1 and β_3) cannot be simultaneously alternative best replies. Alltogether one finds the following equilibria of $G_2(\eta)$ close to (2, 2):

$$\begin{aligned}
&\langle (1 - 3\delta_3, 2\delta_3, \delta_3), (\tfrac{2}{3} - \varepsilon_3, \tfrac{1}{3}, \varepsilon_3) \rangle && \text{if} \quad \delta_2 < 2\delta_3, \\
&\langle (1 - 3\delta_2/2, \delta_2, \delta_2/2), (\tfrac{1}{3} - \varepsilon_3, \tfrac{2}{3}, \varepsilon_3) \rangle && \text{if} \quad \delta_2 > 2\delta_3, \\
&\langle (1 - 3\delta_3/2, \delta_3/2, \delta_3), (\varepsilon_1, \tfrac{1}{3}, \tfrac{2}{3} - \varepsilon_1) \rangle && \text{if} \quad \delta_2 < \delta_3/2, \\
&\langle (1 - 3\delta_2, \delta_2, 2\delta_2), (\varepsilon_1, \tfrac{2}{3}, \tfrac{1}{3} - \varepsilon_1) \rangle && \text{if} \quad \delta_2 > \delta_3/2.
\end{aligned}$$

(If $\delta_2 = 2\delta_3$ then both the first and second strategy pair are equilibria of $G_2(\eta)$, as well as any convex combination of these.) Since every perturbed game $G_2(\eta)$ with η small has an equilibrium close to (2, 2), this equilibrium outcome is indeed stable in G and hence in Γ_2. However, the equilibria of $G_2(\eta)$ are difficult to interpret in Γ_2. Namely consider the first equilibrium and note that, as $\eta \to 0$, the strategy pair induced in the subgame of Γ_2 converges to $\langle (\frac{2}{3}, \frac{1}{3}), (\frac{2}{3}, \frac{1}{3}, 0) \rangle$ and this is not an equilibrium of the subgame! The same property holds for all other equilibria: No equilibrium of $G_2(\eta)$ induces (in the limit) a subgame perfect equilibrium in Γ_2, and this is the reason for why stability does not conform with forward induction in this example.

4. The Importance of Sunk Cost

In this section it is shown that the conventional argument that sunk costs are irrelevant is incorrect whenever there is strategic uncertainty: The amount of costs that is sunk may determine which equilibrium of the strategic interaction is more focal, hence, these costs may determine which equilibrium is actually played. In addition we will show that strategic reasons may force players to sink costs.

Consider again BS from Fig. 1a, but now assume that, before playing this game, player 1 has the opportunity to burn a certain amount of money x. Player 2 does not have this opportunity, but this player can see whether 1 chose to burn the money and this is known to player 1; hence, when playing BS it will be common knowledge whether or not player 1 burnt money. Note that BS can be interpreted as a "battle for market shares game" and that x can be viewed as the amount spent on advertising. The question then is whether 1 will advertise and who will get the bulk of the market.

For the moment, assume x is given exogenously with $x > \frac{1}{4}$. Then the reduced normal form is given in Fig. 5. Player 1 can guarantee a payoff $\frac{3}{4}$ by not burning money and by playing his maxmin strategy $1/4s + 3/4w$ in BS; hence, xw (burning x followed by playing weak) is a dominated strategy for this player, so that player 2 should conclude that 1 plays s after he has burnt x. Consequently, after x player 2 should respond with w and player 1 is "guaranteed" $3 - x$ if he chooses x (i.e., he will get this amount if player 2 subscribes to a theory that says that dominated strategies are irrational). However, this implies that ow (not burning money and then playing weak) is dominated if $x < 2$, so that, in this case, player 2 should

	ss	sw	ws	ww
os	0, 0	0, 0	3, 1	3, 1
ow	1, 3	1, 3	0, 0	0, 0
xs	$-x$, 0	$3 - x$, 1	$-x$, 0	$3 - x$, 1
xw	$1 - x$, 3	$-x$, 0	$1 - x$, 3	$-x$, 0

FIG. 5. Reduced normal form.

conclude that 1 will play s even if he does not burn money. Consequently, if $\frac{1}{4} < x < 2$ there is only 1 outcome that survives iterated elimination of dominated strategies: Player 1 does not burn money but he nevertheless receives his most preferred equilibrium. The availability of an additional option, although not used, has drastic consequences for the solution and the possibility to sink cost is relevant.

Note that above we just used elimination of dominated strategies so that the same result would have been obtained by applying "rationalizability" (Bernheim [3], Pearce [15]) or "dominance solvability" (Moulin [13]). Furthermore, if $1 \leqslant x \leqslant 2$ all dominations are actually by pure strategies.

Let us briefly analyse what happens for other values of x. If $x > \frac{9}{4}$, then both xw and xs are dominated (by the maxmin strategy in BS) so that the additional option is fully irrelevant: The game just reduces to BS. If $2 < x < \frac{9}{4}$, then xw is dominated and player 2 should respond to x with w, so that the normal form reduces to a 2×3 game. Clearly the strict equilibria (3, 1) and (1, 3) are now stable (as singletons), but also the component $(3 - x, 1)$ is stable, intuitively because player 2 can never know whether 1 wants to continue with (3, 1) or with (1, 3) if he chooses not to burn money (cf. Sect. 3). If $0 < x < \frac{1}{4}$ the game of Fig. 5 does not admit dominated (pure) strategies. In this case, the game admits five subgame perfect equilibrium paths: not burning money followed by any of the equilibria of BS (notation $(os, w\cdot)$, $(ow, s\cdot)$, $(om, m\cdot)$) and burning x followed by (s, w) or (w, s) (notation $(xs, \cdot w)$, $(xw, \cdot s)$). We claim that exactly three of these are stable. This stability property clearly holds for (os, w) as player 1 receives his maximal payoff and player 2 cannot signal a deviation. It is also clear from Fig. 5 that $(ow, s\cdot)$ is not stable: xw is an inferior response against the set of equilibria supporting this path so that player 2 should conclude that 1 plays s after a deviation but then player 1 gains from burning x. Exactly the same argument shows that $(xs, \cdot w)$ is not stable. To verify the stability of $(om, m\cdot)$ and of $(xw, s\cdot)$ remains and this can be done by a direct computation where we note that the basic reason for their stability is that a deviation by player 1 cannot be given an unambiguous meaning. The following table summarizes our findings:

Case	Stable paths
$0 < x < \frac{1}{4}$	$(os, w\cdot)$, $(om, m\cdot)$, $(xw, \cdot s)$
$\frac{1}{4} < x < 2$	$(os, w\cdot)$
$2 < x < 2\frac{1}{4}$	$(os, w\cdot)$, $(ow, s\cdot)$, $(xs, \cdot w)$
$x > 2\frac{1}{4}$	$(os, w\cdot)$, $(ow, s\cdot)$, $(om, m\cdot)$

Before turning to a more symmetric version of this game, let us note that if player 1 can choose his level of advertising x from a (finite) set X and

if player 2 can observe the level chosen by 1, then as long as $0 \in X$ and $(\frac{1}{4}, 2) \cap X \neq \varnothing$, there is a unique equilibrium outcome that survives iterated elimination of dominated strategies. Player 1 sets $x = 0$ and chooses s, player 2 responds with w. So, if only one firm has the opportunity to advertise, it will not use this option but it will nevertheless receive the larger market share. (The proof follows the argument given at the beginning of this section.)[10]

Finally, consider the game in which both players can throw away money. Specifically, let X be a finite set with $0 \in X$ and $(\frac{1}{4}, 2) \cap X \neq \varnothing$ and consider the following 2-stage game:

(i) Simultaneously the players choose x_1 (resp. x_2) from X,

(ii) Being informed about (x_1, x_2), the players play BS.

We claim that in any stable equilibrium of this game, money is burnt with positive probability. To prove this claim, note that there are only three Nash equilibrium paths in which money is burnt with probability 0, viz. players have to continue with either (s, w), (w, s), or (m, m) in such a path. Suppose players continue with (s, w) so that player 2's equilibrium payoff is 1. Hence, as long as player 1 follows the path, player 2 can guarantee the payoff 1 by choosing a strategy $(0, f_2)$ with $f_2(0) = w$ (we use $f_2(x)$ to denote player 2's reaction if 1 chooses x in round 1), from which it follows that any strategy (x^*, f_2) with $0 < x^* < 2$ and $f_2(0) = w$ is an inferior response against this component. After having deleted these inferior strategies, the strategies $(0, f_1)$ with $f_1(x^*) = s$ of player 1 become dominated. Consider the reduced game in which these strategies have been eliminated also. In this game player 2 (by playing (x^*, f_2^*) with $f_2^*(0) = s$) is guaranteed a payoff $3 - x^*$ as long as player 1 plays any strategy $(0, f_1)$, i.e., as long as player 1 plays any strategy supporting the original path. However, as $x^* < 2$ this implies that none of the equilibria of the component that we started with remains an equilibrium in the reduced game and Proposition A allows us to conclude that this component is not stable. (Note that just using elimination of dominated strategies would not allow this conclusion.)

Exactly the same argument establishes the instability of the path in which players continue with the mixed equilibrium after $(0, 0)$ in round 1 (however, this time one needs $\frac{1}{4} < x^* < 2$). Hence, we conclude that strategic stability requires both players to throw away money and that all stable equilibria are inefficient.

[10] This observation has been generalised in Ben-Porath and Dekel [2] who showed that in 2-person games, in which there exists a unique Pareto efficient outcome, only this outcome survives iterated elimination of dominated strategies of the extended game in which one player can throw away arbitrary amounts of money.

5. Repeated Games

In this section we wish to indicate that the theory of repeated games has to be modified in an essential fashion if one accepts the concept of strategic stability and/or the idea of forward induction. Specifically, we wish to show that the equilibria used in the proof of the (perfect) Folk Theorem (Benoit and Krishna [1]) fail to be stable. Even though our analysis is fairly superficial and is restricted to two-fold repetitions, it may highlight the following startling results:

(5.1) paths composed of stable equilibria in the one-shot game need not be stable in the repeated game;

(5.2) to verifiy stability one does not only have to worry about profitable one-shot deviations—also deviations resulting in an immediate loss may be problematic;

(5.3) threats to revert to the worst stable equilibrium (i.e., optimal punishments) are typically not credible (according to stability); and

(5.4) the restriction to pure strategies is not justified as there may be no stable equilibrium within this class.

Again consider BS of Fig. 1a but now assume that this game is repeated twice with both players being informed of the outcome at stage 1 before moving at stage 2. A pure (reduced) normal form strategy for this game is a triple $(\alpha, \beta\gamma)$, where α is the choice at stage 1 ($\alpha = s, w$) and β (resp. γ) is the reaction at stage 2 in case the opponent has chosen s (resp. w) at stage 1. Hence, the reduced normal form is an 8×8 bimatrix game. Consider the path in which the players choose (s, w) at both stages. This path clearly is subgame perfect (the threat to continue with (s, w) after every deviation deters deviations) but we claim that it is not stable. Heuristically, one may demonstrate this claim as follows. Suppose players have agreed to play this path but player 2 nevertheless deviates in round 1, hence, this player has payoff 0 in the first round. Now, observing this deviation, it does not make sense for player 1 to think that 2 will play w at stage 2 (in this case, 2's total payoff could maximally be 1 which is less than what this player was "guaranteed" in equilibrium). However, by playing s, player 2 could get more and, furthermore, if player 1 indeed interprets the deviation as a signal that 2 will play s and chooses his best response, then 2 does indeed gain by deviating. Hence, there is only one "rational" inference that 1 can draw from a deviation of 2 and if 1 responds optimally then 2 gains, so that the path is not stable.

More formally, one argues as follows. Consider the equilibrium component of the repeated game in which (3, 1) is played twice. Obviously, the strategies $(w, \cdot\cdot)$ in which player 1 starts with w, as well as the strategies

$(s, \cdot w)$, in which player 1 unilaterally deviates in round 2, are inferior responses against this component as they are sure to yield a payoff less than 6. Hence, for player 1, only (s, ss) and (s, ws) are non-inferior. Similarly, strategies in which player 2 only deviates in round 2 are inferior. Furthermore, given that player 1 always chooses s in round 1, the strategies $(\cdot, \cdot w)$ and $(\cdot, \cdot s)$ are equivalent for player 2, so that the 8×8 normal form can be reduced to the following 2×3 game of Fig. 6.

In this reduced game $(s, w \cdot)$ is an inferior response (against the equilibria with payoff (6, 2)), and after having deleted this, (s, ss) becomes dominated. Hence, iterated elimination reduces the game to (1, 3) which is not in our component, so that (Proposition A), the path in which (3, 1) is played twice is not stable. This verifies our claims (5.1) and (5.2).

Clearly, in the repetition of BS, also playing (1, 3) twice is not stable. However, alternating between (3, 1) and (1, 3) is stable, the (intuitive) reason being that if a player unilaterally deviates from this path he is always sure to get less than he was guaranteed in equilibrium. Hence, any unilateral deviation is inferior and one cannot attach a "rational" meaning to any of them. Furthermore, paths in which players start with the completely mixed equilibrium and then play any equilibrium of BS in round 2 (possibly dependent on the outcome of round 1) are also stable as one cannot detect any deviations in round 1 (and one always continues with a stable equilibrium in round 2). Note, however, that stability of such paths can be destroyed by adding a strictly dominated strategy to BS: If player 1 has a third action available that always yields -1, then playing the mixed equilibrium twice is not stable as player 2 can deduce that 1 will play s after a deviation to this third strategy in the first round. There also exist stable paths that do not consist of playing a one-shot equilibrium at every stage. For example, the players can randomize at stage 1 (each choosing s with probability $\frac{1}{2}$) while continuing with (s, w) after (w, s) and with (w, s) in all other situations. (This equilibrium yields both players a total payoff of $\frac{5}{2}$.) Computing the set of all stable equilibrium paths seems to be laborious, however, even in this most simple conceivable case, and games with a larger horizon should definitely be handled by a computer.

	$(w, w\cdot)$	$(s, w\cdot)$	$(s, s\cdot)$
(s, ss)	6, 2	3, 1	0, 0
(s, ws)	6, 2	0, 0	1, 3

FIG. 6. Alternative proposal reduced game of repeated BS.

In the above example the only stable paths that are in pure strategies involve alternation between the players' most preferred equilibria. The following example may demonstrate that this is not a general property, however.

Let $a<3$ so that the game of Fig. 7 has three stable equilibria, viz. (s, w), (w, s), and a completely mixed one with value $3/(4-a)$. Consider the two-fold repetition of this game. It is easily seen that if $a<2$, then any subgame perfect equilibrium path in which at least one player does not randomize at stage 1 must start with (s, w) or with (w, s) in the first round, hence, there are at most six such paths. Exactly the same reasoning as the one leading to Fig. 6 shows that (s, w) followed by either (s, w) or the mixed equilibrium is not stable (player 1 should conclude that 2 will play s after a deviation), so that, of all the paths not involving randomization at stage 1, only alternation between $(3, 1)$ and $(1, 3)$ is possibly stable. However, also these paths are not stable for $1<a<2$: If player 2 deviates from the path (w, s), (s, w) at stage 1, then player 1 should conclude that player 2 will play s at $t=2$ (since $a<2$ playing w twice is inferior), hence 1 should respond to the deviation by choosing w, but then player 2's deviation is profitable (since $a>1$). (The formal argument is exactly as for BS given above.) We can conclude that, if $1<a<2$, all stable paths start with both players randomizing, which verifies our claim (5.4), establishing inter alia the inefficiency of all stable equilibria in this twofold repetition.

In independent recent work, Osborne [14] has investigated the stability of pure equilibrium paths in coordination games. He formulates a "criterion of immunity against a convincing deviation" (which basically is a translation of the Cho and Kreps [6] "intuitive criterion" to the repeated game context) and he shows a pure path that is not immune against such a deviation cannot be stable. The above example clearly demonstrates that the restriction to pure paths is undesirable. The same example may also show that Osborne's criterion may not exclude all unstable equilibria. Take $2<a<\frac{5}{2}$. Then the path (s, w), (s, w) is unstable (apply Proposition A(ii)), but it is immune against any convincing deviation.

	s	w
s	0, 0	3, 1
w	1, 3	a, a

FIG. 7. Alternative proposal modified BS.

	c	s	w
c	6, 6	0, 8	0, 0
s	8, 0	4, 4	0, 1
w	0, 0	1, 0	1, 1

FIG. 8. Alternative proposal modified prisoners' dilemma.

Up to now we only considered unprofitable one-shot deviations. In the concluding example (a modified prisoners' dilemma) we show that also profitable one-shot deviations may pose a test for stability. This example also demonstrates that the threat to revert to the worst stable equilibrium continuation need not be stable, hence, it verifies (5.3). (Recall that similar threats suffice to guarantee subgame perfectness.)

The game of Fig. 8 has three equilibria with respective payoffs (4, 4), (1, 1), and (1, 1) (payoffs could be perturbed as to make the game generic). The twofold repetition allows $(c, c)(s, s)$ as a subgame perfect equilibrium path (after a deviation players continue with (w, w)), but this path is not stable. If player 1 unilaterally deviates from this path in the first round, then player 2 knows that in the second round this player will not continue with c (as c is strictly dominated), nor with w (as all such strategies are inferior against the component, they yield at most 9). Hence, player 2 should conclude that 1 continues with s, but then he is forced to play s and player 1 gains from deviating.

APPENDIX: REVIEW OF DEFINITIONS

A *normal form* game is a $2n$-tuple $G = (A_1, ..., A_n, g_1, ..., g_n)$, where A_i is a finite nonempty set and $g_i: A \to \mathbb{R}$, where $A = X_i A_i$. We use S_i to denote the set of mixed strategies of player i and $S = X_i S_i$. The probability that s_i assigns to a_i is $s_i(a_i)$ and $s(a) := X_i s_i(a_i)$. The expected payoff is $g_i(s) = \sum_a s(a)\, g_i(a)$. The *reduced normal form* of G is the normal form that results after all pure strategies a_i have been deleted for which there exists some s_i with $s_i(a_i) = 0$ and $g(s' \backslash s_i) = g(s' \backslash a_i)$ for all $s' \in S$. (As usual $s' \backslash s_i = (s'_1, ..., s'_{i-1}, s_i, s'_{i+1}, ..., s'_n)$ and $g = (g_1, ..., g_n)$.) Given a game Γ in *extensive form*, the normal form associated with Γ is defined as usual (see Selten [18]). Note that a strategy in the reduced normal form of Γ does

not tell what player i should do at an information set that cannot be reached as long as player i sticks to his strategy.

s_i is a best reply against s' if $g_i(s' \backslash s_i) \geqslant g_i(s' \backslash s_i'')$ for all s_i'' and s is a *Nash equilibrium* of G if s_i is a best reply against s for all i. The set of Nash equilibria (which is nonempty) consists of finitely many connected *components* and if G results from a generic extensive game then all points in the same component generate the same *outcome*; i.e., they yield the same probability distribution over the endpoints of the tree (Kohlberg and Mertens [9, Appendices B, C]).

Let $\eta = (\eta_1, ..., \eta_n)$, where $\eta_i: A_i \to \mathbb{R}_{++}$ with $\sum_{a_i} \eta_i(a_i) < 1$ for all i. The perturbed game $G(\eta)$ results from the normal form G by restricting each player to those completely mixed strategies satisfying $s_i(a_i) \geqslant \eta_i(a_i)$ for all a_i. A strategy combination s satisfying these restrictions is an equilibrium of $G(\eta)$ if only best responses are chosen with more than minimum probability (i.e., if $s_i(a_i) > \eta_i(a_i)$, then $g_i(s \backslash a_i) \geqslant g_i(s \backslash a_i')$ for all a_i'). A set of equilibria E of G is called *stable* if it is a minimal set with the property that for any sufficiently small η there exists an equilibrium of $G(\eta)$ close to E. With a small abuse of terminology, an equilibrium component is called stable if it contains a stable set and an equilibrium outcome (resp. payoff) of a generic extensive form game is said to be stable if there exists a stable component of the associated reduced normal form that generates this outcome (resp. payoff). Kohlberg and Mertens [9] showed that every game admits a stable component, hence, generic extensive games have stable outcomes and stable payoffs.

Strategy s_i' is said to be *dominated* if there exists s_i'' with $g_i(s \backslash s_i'') \geqslant g_i(s \backslash s_i')$ for all s with at least one inequality being strict. s_i'' is said to be an *inferior response* against a set of equilibria E if there does not exist an element of E against which s_i' is a best reply. Note that a dominated strategy need not be inferior and that an inferior strategy need not be dominated.

References

1. J.-P. Benoit and V. Krishna, Finitely repeated games, *Econometrica* **53** (1985), 905–922.
2. E. Ben-Porath and E. Dekel, "Coordination and the Potential for Self Sacrifice," mimeo, Dept. of Economics, Univ. of California, Berkeley, November 1987.
3. B. D. Bernheim, Rationalizable strategic behavior, *Econometrica* **52** (1984), 1007–1029.
4. K. G. Binmore, Equilibria in extensive games *Econ. J.* **95** (1984), 51–59.
5. K. G. Binmore, "Remodeled Rational Players," mimeo, London School of Economics, April 1987.
6. I. K. Cho and D. Kreps, Signaling games and stable equilibria *Quart. J. Econ.* **102** (1987), 179–221.
7. J. C. Harsanyi and R. Selten, "A General Theory of Equilibrium Selection in Games," MIT Press, Cambridge, MA, 1988.

8. E. KALAI AND D. SAMET, Persistent equilibria in strategic games, *Int. J. Game Theory* **13** (1984), 129–144.
9. E. KOHLBERG AND J.-F. MERTENS, On the strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1037.
10. D. FUDENBERG, D. KREPS, AND D. LEVINE, On the robustness of equilibrium refinements, *J. Econ. Theory* **44** (1988), 354–380.
11. D. KREPS AND R. WILSON, Sequential equilibria, *Econometrica* **50** (1982), 863–894.
12. J.-F. MERTENS, "Ordinality in Noncooperative Games," Core Discussion Paper 8728, Université Catholique de Louvain, 1987.
13. H. MOULIN, Dominance solvable voting schemes, *Econometrica* **47** (1979), 1337–1351.
14. M. OSBORNE, "Signaling, Forward Induction and stability in Finitely Repeated Games," mimeo, Dept. of Economics, McMaster University, November 1987.
15. D. PEARCE, Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52** (1984), 1029–1051.
16. J. P. PONSSARD, "Interactive Plans for Extensive Games," mimeo, Laboratoire d'Econometrie de l'Ecole Polytechnique, Paris, January 1988.
17. T. C. SCHELLING, "The Strategy of Conflict," Harvard Univ, Press, Cambridge, MA, 1960.
18. R. SELTEN, Reexamination of the Perfectness concept for equilibrium points in extensive games, *Int. J. Game Theory* **4** (1975), 25–55.
19. E. VAN DAMME, "Stable Equilibria and Forward Induction," D.P. A-128, SFB 303, Bonn, August 1987.
20. J. VON NEUMANN AND O. MORGENSTERN, "Theory of games and economic behavior," Princeton Univ. press, Princeton, NJ, 1948.
21. J. WEIBULL, "Refinements of Subgame Perfection—Without Trembles" D.P. A-146, SFB 303, Bonn, January 1988.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium